

Master 2 Biologie-Informatique/ Bioinformatique



Etude de la fonction et des mécanismes d'évolution des séquences répétées centromériques chez les Primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Muséum national d'Histoire naturelle, 43 rue Cuvier 75005 PARIS



Remerciements

Merci à Namrod pour toute la partie sur la bibliographie. Retrouvez ses questions FAQ qui ont permis la rédaction de cette partie.

Merci à f-leb, LittleWhite et Metalman pour leurs conseils et la relecture. Merci à ced et jacques_jean pour la correction orthographique et typographique.

Table des matières

Remerciements	1
1 Introduction	1
1.1 Le centromère	1
1.2 L'ADN α -satellites	1
1.3 Le sujet de stage	2
2 Matériel et méthode	3
2.1 Choix des espèces	3
2.2 Méthode de classification	3
2.2.1 Principe	3
2.2.2 Répartition itérative	4
2.2.3 Double-validation d'un sous-groupe	4
2.3 Analyse des séquences	4
3 Résultat	5
3.1 Caractérisation intraspécifique des familles	5
3.1.1 Identification des familles	5
3.1.2 Motifs potentiellement fonctionnels	9
3.1.3 Similarité entre familles	11
3.2 Comparaison inter-espèce	11
3.2.1 Répartition des super-familles	11
3.2.2 Une mosaïque de familles C2	12
3.2.3 Origine du motif pK β	13
4 Discussion	14
5 Conclusion	14

1 Introduction

1.1 Le centromère

Le centromère est une structure chromatinienne caractérisé par la présence de CENP-A. Cette protéine, très conservée au cours de l'évolution, est un variant de l'histone H3. Son rôle est de fixer la position du kinétochore par un mécanisme encore peu connu. En effet, le centromère est le site d'assemblage du kinétochore, un ensemble d'ADN et de protéines. Il permet l'attachement du fuseau mitotique pour la ségrégation des chromosomes durant la division cellulaire chez les eucaryotes. Le centromère et les protéines impliquées sont relativement bien conservés. Au contraire, l'ADN sous-jacent est très diversifié et l'organisation varie d'un taxon à l'autre. Cependant, une caractéristique commune est retrouvée chez toute les espèces : de l'ADN centromérique répété en tandem nommé ADN satellite. Ces répétitions sont issues d'événements d'amplification, tels les crossovers inégaux, la conversion de gènes, les cercles roulants ou la transposition de séquences.[Malik and Henikoff, 2002 ; Plohl et al. 2012] Ces séquences représentent 5% du génome. Les répétitions s'étendent de 7pb à 3,2kb avec des séquences de 145-180kb le plus souvent.

1.2 L'ADN α -satellites

L'ADN satellite chez les Primates est connu sous le nom d' α -satellite. Ces séquences centromériques répétées en tandem sont riches en AT.

Des études chez l'Homme propose un modèle évolutif. La répartition des α -satellites suivrait une répartition spécifique selon l'âge des familles. Les familles les plus jeunes s'insèrent au cœur du centromère, repoussant les familles les plus anciennes jusqu'aux regions voisines, appelé péri-centromère.

Un monomère a une longueur de 171pb et il peut être répété des milliers de fois. Les monomères peuvent être répartis en famille selon leur similarité, les séquences ayant un taux d'identité supérieur à 70%. Ces séquences ont soit une organisation monomérique soit une organisation en répétition d'ordre supérieur (Fig. 1). Dans le premier cas, les séquences d'une même famille sont répétées en tandem. Dans le deuxième cas, une suite de monomères appartenant à différentes familles forme une unité, qui elle est répétée en tandem.

Ces séquences peuvent avoir un site de liaison à la protéine centromérique CENP-B un mo-

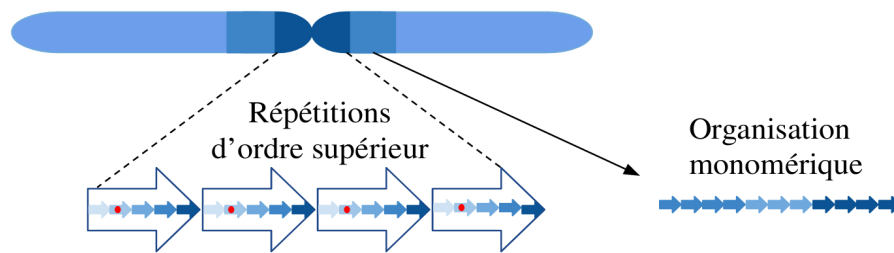


FIGURE 1 – **Organisation spatiale des α -satellites** : Le coeur du centromère (bleu foncé) est organisé en répétition d'ordre supérieur. Le péricentromère (bleu clair) a une organisation monomérique. Un monomère d'une même famille est représenté par une petite flèche de même couleur. Les points rouges représentent les sites de fixation à CENP-B ou pJ α .

tif spécifique de 17pb. Cette protéine, qui reconnaît et se fixe sur l'ADN, serait présente chez de nombreuses familles de Primates. La protéine pJ α , une protéine peu caractérisée, reconnaît un motif qui remplace celui de CENP-B.

Les α -satellites ont essentiellement été étudiées chez l'homme. Modèle évolutif avec les centromères en expansion. Une hypothèse concernant l'âge des séquences découle de ces recherches : les séquences les plus récentes apparaissent au coeur du centromères, déplaçant les plus anciennes au péricentromère. D'autres études chez le gorille ont été faites. Le rôle des α -satellites est encore mal connu.

1.3 Le sujet de stage

Peu d'informations sur les α -satellites figurent chez les autres espèces de primates, et aucune relation inter-espèce n'a été réalisée. Les études se basant sur un séquençage haut-débit est appliqué chez l'homme (cité ci-dessus) et chez le Gorille. [compléter] L'équipe d'accueil de mon stage "ADN répété, Chromatine, Evolution" ou ARChE, a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α -satellites chez deux espèces : *Cercopithecus solatus* et *Cercopithecus pogonias*.

Les méthodes basées sur l'alignement et la phylogénie sont très limitées pour étudier ces séquences. Elles ne permettent pas de traiter des jeux de données conséquents, or un monomère peut avoir des milliers de copies dans un seul génome. De plus, ces méthodes non-objectives ne permettent pas de faire des comparaisons entre espèces.

Pour remédier à ce problème, une méthode de classification automatisée des α -satellites a été implémentée en R en 2016, puis améliorée en 2017 en Python dans le laboratoire. Ce programme permet de traiter des centaines de milliers de séquences, quelque soit le nombre ou

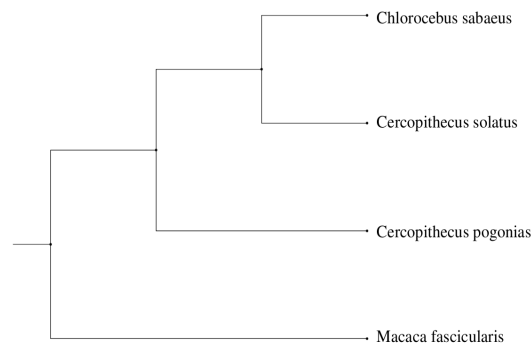


FIGURE 2 – **Arbre phylogénétique des espèces analysées.**[1]

la taille des familles. De plus, cette méthode est objective et peut être appliquée à plusieurs espèces, permettant ainsi une comparaison inter-espèce des familles α -satellites.

L'objectif de ce stage est de comprendre la fonction des α -satellites, notamment en caractérisant les familles issues de cette classification chez quatre espèces proches de primates. Parmi ces espèces, les deux Cercopithèques séquencés dans le laboratoire permettront d'avoir un avis objectif sur la méthode de classification. Dans un deuxième temps, les mécanismes d'évolution pourront être déduits à partir d'une comparaison inter-espèce révélant les différences et les familles communes.

2 Matériel et méthode

2.1 Choix des espèces

Les critères de sélection dépendent de la disponibilité des séquences de qualité. Deux espèces du laboratoire sont choisies, le *C. solatus* et *C. pogonias*, et deux espèces proches (Fig. 2), le *Macaca fascicularis* et le *Chlorocebus sabaeus*.

2.2 Méthode de classification

2.2.1 Principe

Cette méthode [2] répartit des séquences α -satellites en familles selon la similarité. La classification est hiérarchique dichotomique. Au départ, une table contenant les fréquences des 5-mers est calculée pour chaque monomère. Ensuite une boucle itérative est exécutée pour séparer les séquences en groupes tant que les nouveaux groupes formés sont divisibles.

2.2.2 Répartition itérative

Une Analyse en Composante Principale (ACP) est effectuée sur la table des fréquences des 5-mers afin de réduire les dimensions du jeu de données et d'obtenir des variables indépendantes. Des distances euclidiennes sont calculées entre toutes les paires de séquence dans l'espace défini par les premières composantes de l'ACP.

A partir du calcul de distance, les séquences sont séparées en deux classes en utilisant la classification hiérarchique basée sur la méthode de Ward. Cette méthode maximise l'inertie interclasse. La classification hiérarchique fait un usage important de la mémoire. Par conséquent, pour traiter des jeux de données importants de plus de 100 000 séquences, l'Analyse Discriminante Linéaire, une méthode d'apprentissage, est utilisée sur un sous jeu de données formé par de 100 000 séquences tirées aléatoirement, dans ces analyses. Le modèle construit est alors appliqué sur toutes les séquences.

2.2.3 Double-validation d'un sous-groupe

Le premier critère de validation est la taille du sous-groupe. Si un groupe atteint 100 séquences, il n'est pas redivisé. Le deuxième critère de validation s'appuie sur le *matepair*. Ce terme correspond à la proportion de monomères ayant son plus proche voisin dans la même classe, se basant sur les distances euclidiennes calculées auparavant. Des valeurs *matepairs* élevées (proches de 1) indiquent des sous-groupes bien homogènes et séparés validant la classification tandis qu'un seuil *matepair* plus faible (proche de 0) entraîne plus de classes.

Un seuil de *matepair* est fixé à 0.90, pour avoir des groupes homogènes. Si au moins une des valeurs de *matepair* est au-dessous de ce seuil, les sous-groupes sont considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.

2.3 Analyse des séquences

L'alignement des séquences est fait avec muscle [3] et SeaView [?], un éditeur d'alignements multiples. La phylogénie est construite avec la méthode du maximum de vraisemblance (PhyML) [?]. Le modèle F84 est utilisé pour la construction de l'arbre. Le support de branche est aLRT (SH-like). La fréquence d'équilibre nucléotidique est optimisée. Le ratio de transition et de

Espèce	<i>C. solatus</i>	<i>C. pogonias</i>	<i>C. sabaeus</i>	<i>M. fascicularis</i>
Nb. seq. au total	105 529	112 902	29 842	235 535
Nb. fam.	564	132	338	3694
Nb. grandes fam.	12	13	43	114
% seq. ignorées	3.97	1.29	10.89	11.06

TABLE 1 – Résumé du jeu de données et des résultats préliminaires de la classification.

transversion est fixé à 4. Aucun site est considéré comme invariable. Le taux de variation à travers le site est optimisé. Les opérations de recherche d'arbre est NNI et l'arbre de départ est défini avec la méthode de Neighbor-Joining [?] avec une topologie optimisée. Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B (TTCGTTG-GAA[AG]CGGGA), PJ α (TTCCTTTT[CT]CACC[AG]TAG) et pK β (CTATAGGGCCAAAGGAA) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [4] et en autorisant 2 différences au maximum par rapport au consensus.

3 Résultat

3.1 Caractérisation intraspécifique des familles

3.1.1 Identification des familles

A l'issue de la classification, seules les familles ayant plus de 100 séquences, qualifiées de grandes familles, sont conservées pour l'analyse. Le nombre de familles conservées diminue considérablement après élimination des petites familles (<100 séquences) mais ces familles ne représentent que 11% du jeu de données au plus (Tableau 1). Les *C. solatus* et *C. pogonias* ont une dizaine de familles, *C. sabaeus* en a 43, et le *M. fascicularis* en a 114. Malgré le nombre de familles qui diffère d'une espèce à l'autre, la distribution des familles est similaire chez les quatre espèces. Les plus petites familles sont très nombreuses et la fréquence diminue quand la taille augmente (Figure 3). Le *C. solatus* possède une grande famille de 80 000 séquences, *C. pogonias* de 94 000 séquences, *M. fascicularis* possède quatre familles de plus de 10 000 séquences. La plus grande famille du *C. sabaeus* fait 5000 séquences. Ces familles se démarquent des autres familles par leur taille.

Les espèces *C. solatus* et *C. pogonias* sont analysées dans un premier temps pour comparer la classification automatisée avec la classification expérimentale, une méthode visuelle établie d'après une ACP (Fig. 4). Expérimentalement, 6 familles α -satellites ont été déterminées chez les

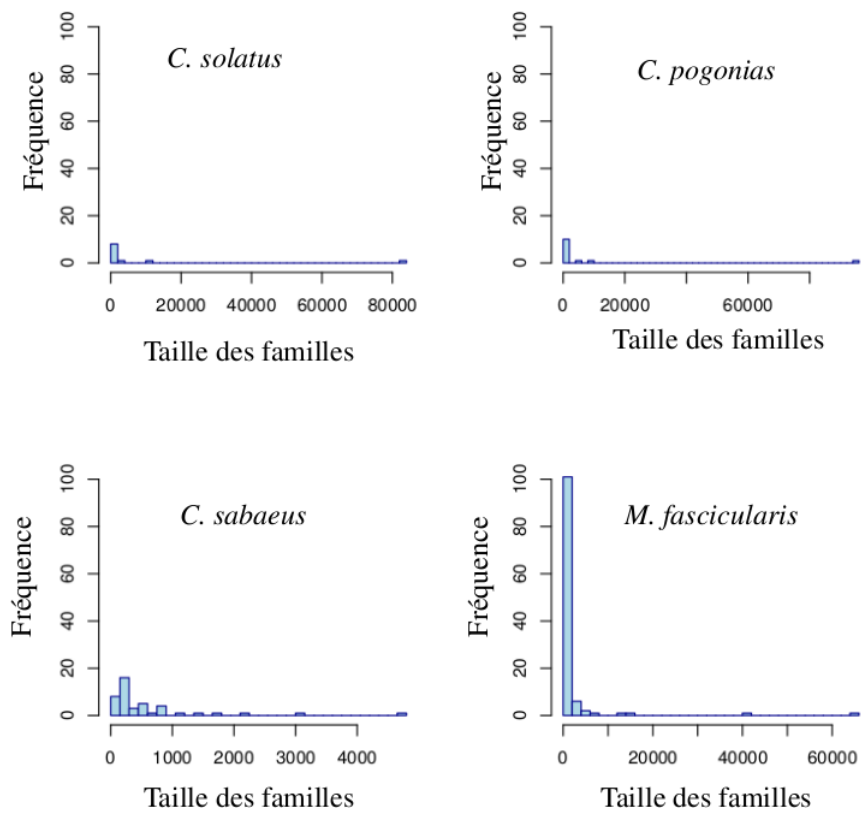


FIGURE 3 – **Distribution des familles en fonction de la taille.**

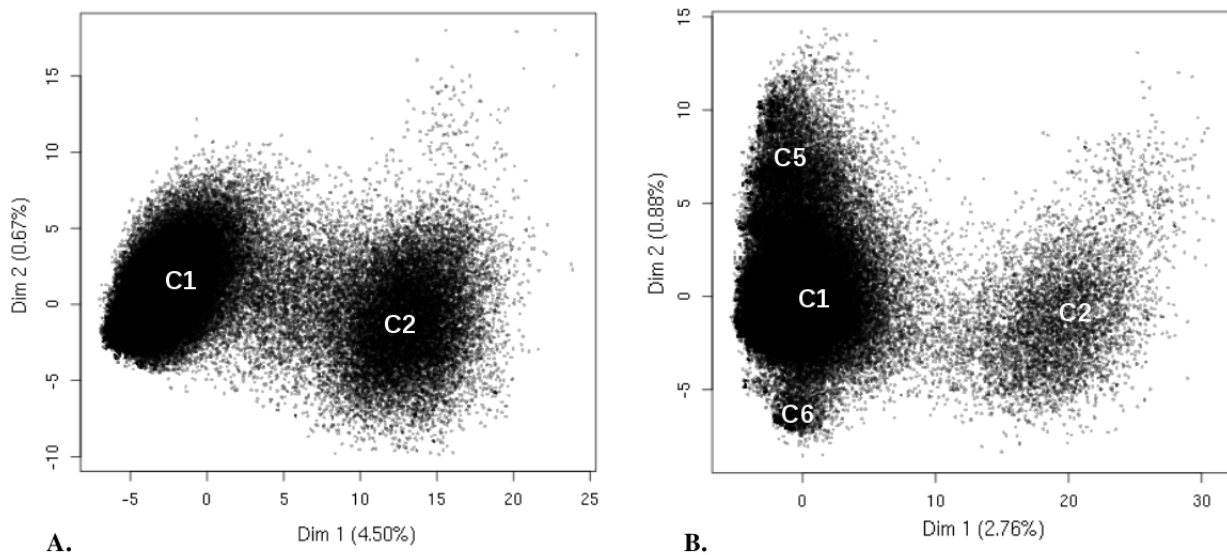


FIGURE 4 – **Caractérisation visuelle des familles α -satellite chez *C. solatus* et *C. pogonias* à partir d'une ACP :** Le nom des familles est indiqué sur les graphiques. Un point représente un monomère. **A.** *C. solatus*. **B.** *C. pogonias*.

Espèces Fam. exp.	<i>C.solatus</i>	<i>C.pogonias</i>
C1	1	10
C2	11	1
C3	(1)	1
C4	(1)	(1)
C5	-	1
C6	-	0

TABLE 2 – **Résumé du tableau de contingence** : Comptage des familles issues de la classification et leur répartition théorique dans les familles expérimentales (C1 à C6). Les valeurs entre parenthèse sont des petites familles (< 100 séquences) qui ne sont pas prises en compte dans le reste des analyses.

Cercopithèques. Ces deux espèces partagent deux grandes familles monomériques, C1 et C2, et deux familles formant un dimère, C3-C4, de l'ordre d'une centaine de séquences chacune. Le *C. pogonias* possède les familles supplémentaires C5 et C6.

Bien que le nombre de grandes familles soit relativement proche entre ces deux espèces, les résultats diffèrent significativement (Tableau 2). Toutes les familles chez le *C. solatus* sont retrouvées : 11 familles forment la famille C2, une famille forme la famille C1 et les familles C3 et C4 sont retrouvées dans des petites familles d'environ 80 séquences chacune. Chez *C. pogonias*, toutes les familles sont retrouvées sauf la famille C6. La famille C1 est répartie en 10 familles, les familles C2, C3 et C5 sont retrouvées entièrement, la famille C4 est également retrouvée sous la forme d'une petite famille de 86 séquences. Chez le *C. solatus*, la famille C2 est divisée en plusieurs familles et la famille C1 est retrouvée dans une seule famille. La situation inverse est retrouvée chez *C. pogonias*.

Pour visualiser cette comparaison, des couleurs sont assignées aux familles issues de la classification automatisée. Ces couleurs sont superposées aux résultats expérimentaux en noir. Chez *C. solatus*, la famille C1 est entièrement retrouvée dans une famille. La famille C2 est répartie en plusieurs familles. Deux familles intermédiaires (rouge et bleue) sont visibles entre la famille C1 (vert) et C2 (orange). Elles ne sont pas distinctes. Une famille supplémentaire (turquoise) se démarque. Pour confirmer cette division de la famille C2, la visualisation de l'ACP des 5-mers est observée en fonction des composantes 1 et 3. Les familles intermédiaires sont toujours confondues, contrairement à la famille turquoise qui forme une famille à part entière. Pour certifier ce fait, l'arbre construit atteste que chaque famille est bien retrouvée, notamment les familles intermédiaires qui forment bien deux familles. Chez *C. pogonias*, les familles C2, C4 et C5 sont bien retrouvées. La famille C6 se fond dans la famille C1 (en vert). La famille C1 est

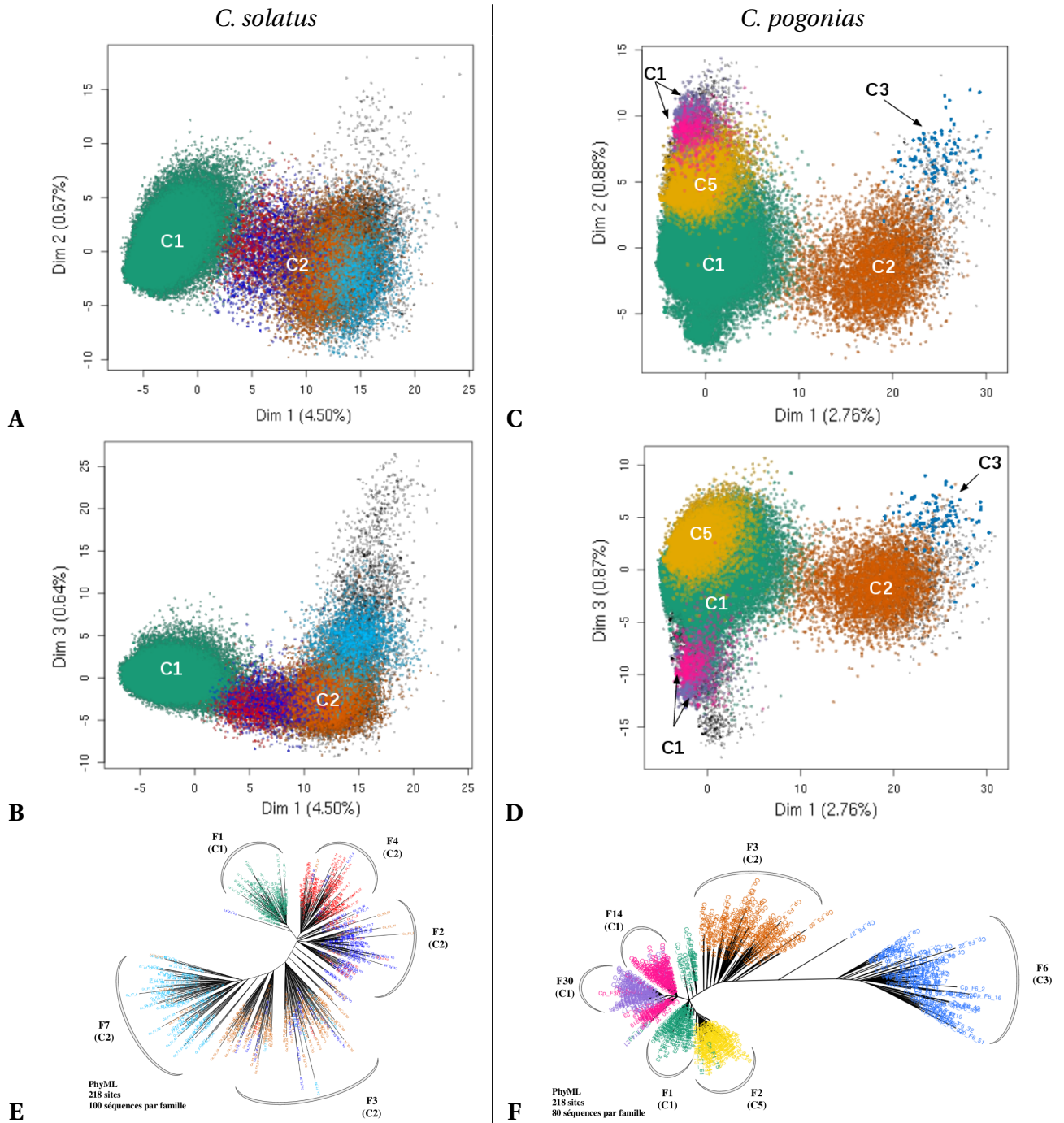


FIGURE 5 – Représentation des plus grandes familles issues de la classification automatisée : Ces familles sont superposées sur les représentations de l'ACP des 5-mers. **A.** Composantes 1 et 2 de l'ACP. Les familles qui correspondraient à C1 sont en vert, C2 en orange, rouge, bleu et turquoise. **B.** Composantes 1 et 3 de l'ACP. **C.** Composantes 1 et 2 de l'ACP. Les familles qui correspondraient à C1 sont en vert, violet et rose, C2 en orange, C4 en bleu clair et C5 en jaune. **C.** Composantes 1 et 3 de l'ACP. **E. et F.** Phylogénie des différentes familles chez *C. solatus* (100 séquences par famille) et *C. pogonias* (80 séquences par famille) respectivement. Les couleurs sont respectivement conservées.

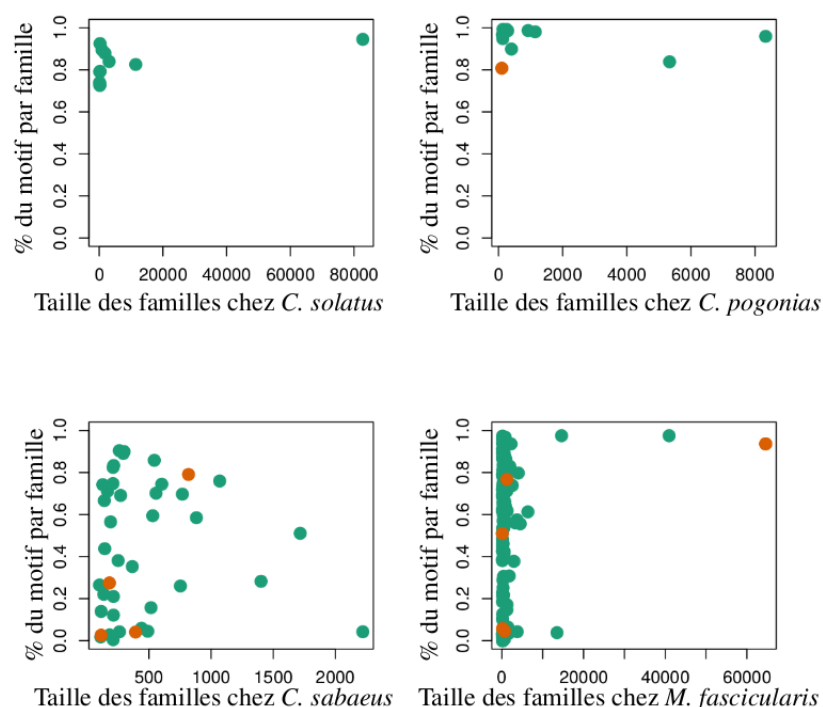


FIGURE 6 – **Présence des motifs CENP-B, pJ α ou pK β par famille** : Le pourcentage de séquences par famille ayant le motif pJ α est en vert, pK β en orange et CENP-B en bleu. Chaque famille est représentée en fonction de sa taille.

divisée en deux familles supplémentaires (rose et violet). La visualisation des composantes 1 et 3 de l'ACP ne permet pas de trancher sur la classification. L'arbre montre que les familles en rose et violet sont très proches.

3.1.2 Motifs potentiellement fonctionnels

La protéine CENP-B est présente chez toutes les espèces, mais les Cercopithèques ne possèdent pas son site de liaison. La protéine pJ α est une protéine peu connue mais dont le site de liaison est déterminé. Le motif pK β est un site commun aux familles n'ayant ni CENP-B ni pJ α . Ces trois motifs sont recherchés pour chaque famille de chaque espèce sur la base des consensus (partie matériel et méthodes) avec 2 mismatches autorisés dans le but de caractériser ces familles. En effet, CENP-B est absent chez *C. solatus* et *C. pogonias*. Le *C. sabaeus* et le *M. fascicularis* n'ont pas ce motif non plus. Par contre plus de 90% des familles chez les quatre espèces ont le motif pJ α mais à des niveaux différents.

La totalité des familles chez *C. solatus* ont le motif, dont 8 à plus de 75%. Chez le *C. pogonias*, 12 familles sur 13 ont le motif à plus de 75%. Par ailleurs, la plus grande famille, C1, ayant plus de 80 000 séquences chez ces deux espèces, se démarque avec un pourcentage à 95%. Le *C. sa-*

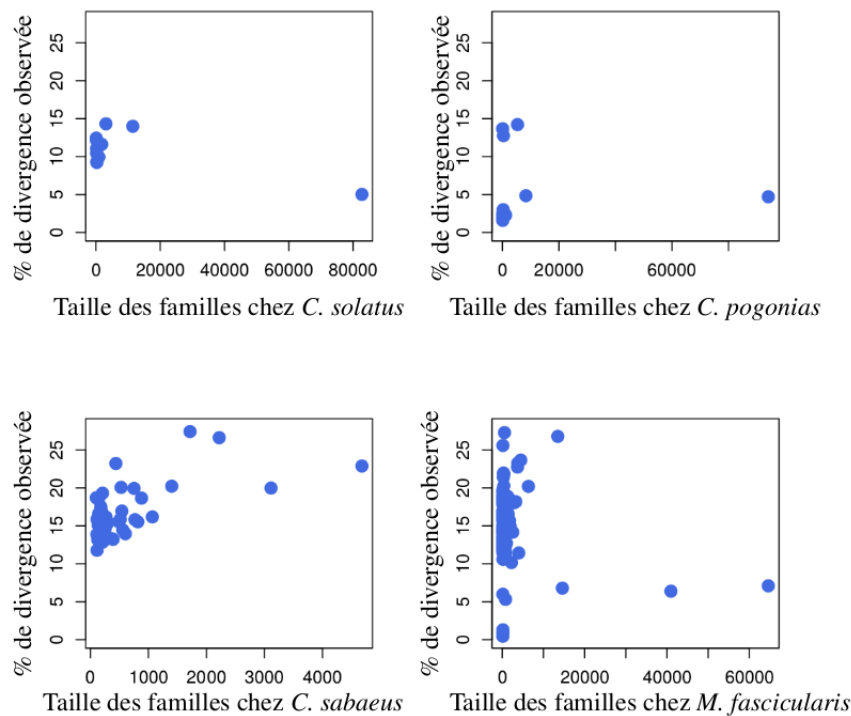


FIGURE 7 – **Pourcentage de divergence observée au sein d’une famille** : Un point correspond à une famille et son pourcentage de divergence observé.

baeus présente 39 familles avec ce motif, dont 7 l’ayant à plus de 75%. Le *M. fascicularis* a des pourcentages pour le motif $pJ\alpha$ qui varie entre 1% et 97%. Cette espèce a 28 familles avec le motif à plus de 75% dont une famille de 40 000 séquences et une autre de 14 000 séquences.

Le motif $pK\beta$ est présent lorsque $pJ\alpha$ est absent de la famille. Il est absent chez *C. solatus*. Seule la famille C3 a ce motif à 80% chez *C. pogonias*. Le *C. sabaeus* a quatre familles avec le motif, dont deux à 79% et l’autre à 29%. Le *M. fascicularis* a cinq familles avec ce motif. La famille ayant le motif à 93% est une grande famille de 64 000 séquences. Les deux autres familles ont le motif à 0.76% et 0.50%.

Les quatre espèces ont des points en commun concernant l’absence du motif CENP-B et quelques familles ayant $pK\beta$. Cependant pour le motif $pJ\alpha$, les *C. solatus* et *C. pogonias* ont des pourcentages relativement proches mais qui diffèrent du *M. fascicularis* et du *C. sabaeus*, dont les valeurs sont intermédiaires. Chaque famille a un motif pour toutes les espèces excepté le *M. fascicularis* qui a 6 familles sans motifs.

3.1.3 Similarité entre familles

Le pourcentage de divergence observée par famille est calculé sur 500 séquences tirées aléatoirement dans une famille pour estimer l'âge de celles-ci. Les *C. solatus* et *C. pogonias* ont des pourcentages relativement faibles ne dépassant pas 15%, le *M. fascicularis* a des valeurs intermédiaires variant de 0.1% à 28%, et *C. sabaesus* a les valeurs les plus élevées allant de 11% à 28%. Les familles qui correspondraient aux familles C1 et C5 ont un pourcentage autour de 5%. Les autres familles chez *C. solatus* ont une moyenne de 10.85%. Les familles de taille moyenne (quelques centaines de séquences) qui correspondraient à la famille C1 ont des pourcentages inférieurs à 3% chez *C. pogonias* et les trois familles restantes ont des valeurs de 13% en moyenne. Les familles du *C. sabaesus* ont en moyenne 16.8% de divergence observée, la taille n'ayant aucun rapport au pourcentage. Le *M. fascicularis* a trois familles avec un pourcentage inférieur à 1% et cinq familles à 6% en moyenne. Les 106 familles restantes ont un pourcentage supérieur à 10%.

Le pourcentage de similarité est semblable chez les *C. solatus* et *C. pogonias*. Le *M. fascicularis* a quelques familles qui ont le même comportement mais il a également beaucoup de familles aux pourcentages très élevés, tandis que la majorité des familles chez le *C. sabaesus* a une grande partie de ses familles avec de grands pourcentages.

3.2 Comparaison inter-espèce

Pour étudier les mécanismes d'évolution des α -satellites, une classification inter-espèce ou "super-classification" (SC) permet de comprendre les différences entre espèces. Pour chaque espèce, 100 séquences par grande famille (> 100 séquences) sont tirées aléatoirement. Un jeu de données de 18 100 séquences est soumis à la classification automatique. A l'issue de cette super-classification, 158 familles sont obtenues au total, dont 90 grandes familles. Les petites familles de moins de 20 séquences, soit 1.76% de ce jeu de données, ne sont pas prises en compte dans l'analyse.

3.2.1 Répartition des super-familles

Parmi les super-familles (SF), 38 ont une taille comprise entre 100 et 20 compris, et trois familles ont une taille strictement supérieure à 800 séquences. Seule une SF *a priori* est commune

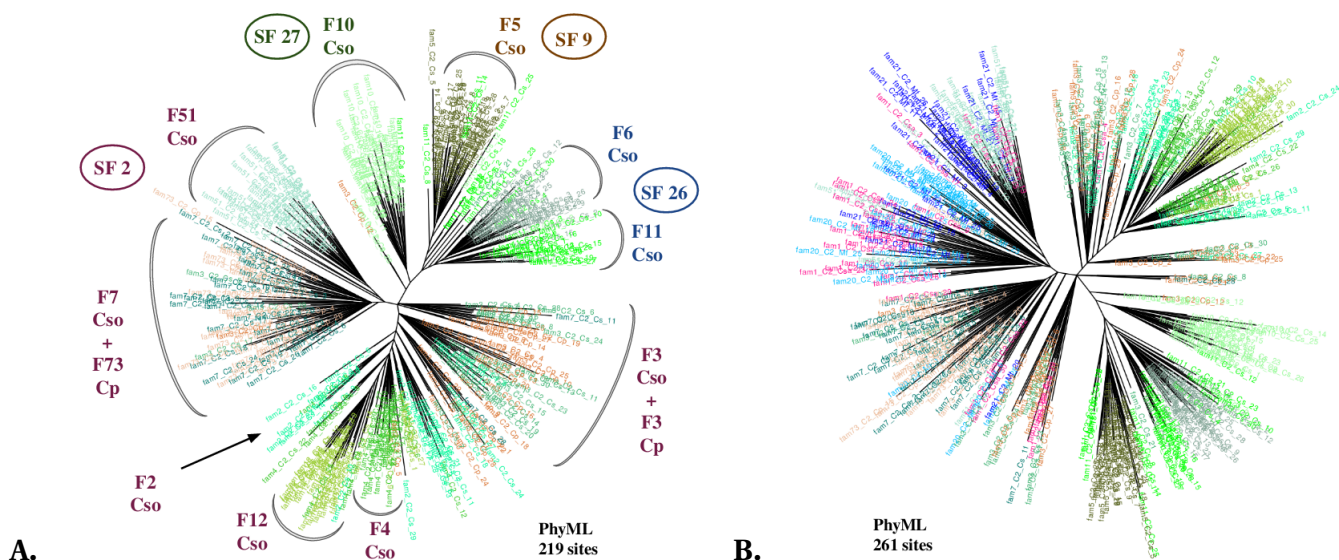


FIGURE 8 – **Phylogénie des différentes familles qui correspondraient à la famille C2 : A.** *C. solatus* (tons verts) et *C. pogonias* (tons orange) **B.** A ces deux espèces sont rajoutés des familles du *M. fascicularis* (tons bleus) et *C. sabeus* (rose).

aux quatre espèces. Elle rassemble 6 familles parmi les 10 familles classées C2 de *C. solatus* et les deux familles également classées C2 de *C. pogonias*, ainsi que deux familles du *M. fascicularis* et une famille du *C. sabeus*. Une partie de la famille annotée C2 serait donc commune aux quatre espèces.

Une SF est partagée entre le *C. pogonias*, le *C. sabeus* et le *M. fascicularis*. Cette SF regroupe des familles ayant le motif pK β et qui correspondrait à la famille C3 commune au *C. pogonias* et au *C. solatus*. Cela signifierait que la famille C3 serait commune à ces quatre espèces également. Certaines SF sont spécifiques aux espèces. Au sein des 8 SF spécifiques du *C. pogonias*, seule l'une d'entre elles regroupe deux familles, le reste étant composé d'une famille. Elles correspondraient aux familles C5, une petite partie de C6 et essentiellement à C1. Les trois SF spécifiques de *C. solatus* seraient équivalentes aux familles C2. Le *C. sabeus* en a 3 et le *M. fascicularis* 27. Le *C. sabeus* et le *M. fascicularis* partagent 38 SF, soit 42% des grandes SF (> 20 séquences), dont la plus grande faisant une taille de 1194 séquences. Une seule SF est uniquement commune à *C. solatus* et *C. pogonias* et elle regroupe les deux plus grandes familles qui seraient du C1.

3.2.2 Une mosaïque de familles C2

La famille C2 fait l'objet d'une séparation en plusieurs familles intéressante. Pour vérifier les résultats de cette SC, 30 séquences de chacune des familles C2 de *C. solatus* et *C. pogonias* sont

"famille C3" en commun, peut-être sous forme de dimère C2-C3. Le deuxième groupe de famille $pK\beta$ est formé de la famille 27 de *C.sabeus* et les familles 47 et 294 du *M. fascicularis* qui ont le motif à 27%, 16% et 41% respectivement. La super-famille 106, regroupant la famille 294, est spécifique de *M. fascicularis*. Les deux familles restantes, qui ont un faible pourcentage du motif, se regroupent pour former la super-famille 115. De plus, ces super-familles sont uniquement constituées de familles $pK\beta$ et ne se mélangent pas aux autres familles. Un arbre composé de toutes les super-familles $pK\beta$ est construit pour voir comment elles s'assemblent. Le jeu de données est construit à partir de 50 séquences par famille $pK\beta$ par espèce tirées aléatoirement (Figure 9) et contiennent la famille C3 de *C. solatus*, dont la famille était trop petite pour être prise en compte. Comme les autres "familles C3", elle se range dans la super-famille 20.

4 Discussion

Les tailles moyennes des consensus sont de 173 nucléotides pour *C. pogonias*, 173 pour *C. solatus*, 173 pour le *C. sabaeus* et X pour le *M. fascicularis*

Les plus grandes familles C2 de *solatus* sont regroupées

5 Conclusion

Références

- [1] Cacheux. Evolutionary history of alpha satellite dna in cercopithecini.
- [2] Florence Jornod. Master's thesis, Paris Diderot, 2016-2017.
- [3] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput, 2004.
- [4] Peter Rice, Ian Longden, and Alan Bleasby. Emboss : the european molecular biology open software suite, 2000.

Résumé

Votre résumé commence ici... ...

Abstract

Abstract begins here... ...